Calculation of likelihood ratios for inference of biological sex from human skeletal remains

Morrison, Geoffrey Stewart; Weber, Philip; Basu, Nabanita; Puch-Solis, Roberto; Randolph-Quinney, Patrick

DOI:
10.1016/j.fsisyn.2021.100202

Publication date:
2021

Forensic Science International: Synergy 3 (2021) 100202

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy

Calculation of likelihood ratios for inference of biological sex from human skeletal remains

Geoffrey Stewart Morrison [a,b,*], Philip Weber [a], Nabanita Basu [a], Roberto Puch-Solis [c], Patrick S. Randolph-Quinney [d,e]

[a] Forensic Data Science Laboratory, Computer Science Department & Aston Institute for Forensic Linguistics, Aston University, Birmingham, UK
[b] Forensic Evaluation Ltd, Birmingham, UK
[c] Leverhulme Research Centre for Forensic Science, University of Dundee, Dundee, UK
[d] Forensic Science Research Group, Department of Applied Sciences, Northumbria University, Newcastle upon Tyne, UK
[e] Department of Human Anatomy and Physiology, University of Johannesburg, South Africa

ARTICLE INFO

Keywords:
Forensic inference and statistics
Forensic anthropology
Likelihood ratio
Sex assessment
Osteometry

ABSTRACT

It is common in forensic anthropology to draw inferences (e.g., inferences with respect to biological sex of human remains) using statistical models applied to anthropometric data. Commonly used models can output posterior probabilities, but a threshold is usually applied in order to obtain a classification. In the forensic-anthropology literature, there is some unease with this "fall-off-the-cliff" approach. Proposals have been made to exclude results that fall within a "zone of uncertainty", e.g., if the posterior probability for "male" is greater than 0.95 then the remains are classified as male, and if the posterior probability for "male" is less than 0.05 then the remains are classified as female, but if the posterior probability for "male" is between 0.05 and 0.95 the remains are not classified as either male or female. In the present paper, we propose what we believe is a simpler solution that is in line with interpretation of evidence in other branches of forensic science: implementation of the likelihood-ratio framework using relevant data, quantitative measurements, and statistical models. Statistical models that can implement this approach are already widely used in forensic anthropology. All that is required are minor modifications in the way those models are used and a change in the way practitioners and researchers think about the meaning of the output of those models. We explain how to calculate likelihood ratios using osteometric data and linear discriminant analysis, quadratic discriminant analysis, and logistic regression models. We also explain how to empirically validate likelihood-ratio models.

1. Introduction

Forensic anthropology is the medico-legal application of biological anthropology. Forensic anthropologists apply to the analysis of human remains detailed knowledge of the development, the morphology, and the normal and abnormal variation of the human body. Analyses are conducted in order to assist legal-decision makers to make decisions with respect to identity of human remains [1-3]. Forensic anthropologists assist in the identification of individuals whose remains are severely decomposed, burned, disrupted, mutilated, or otherwise rendered difficult to recognize, particularly in cases where DNA evidence or odontological evidence are not available. Forensic anthropologists work on investigations related to unexplained natural deaths, accidents, homicide, war crimes, and genocide. They also increasingly work on disaster-victim identification, i.e., investigations related to mass fatality such as occur in building collapses, ship sinkings, and plane crashes.

Forensic anthropologists conduct evaluations with respect to chronological age, biological sex, living stature, and ancestry or population affinity. The analytical methods used can be divided into:

- morphoscopic, i.e., based on visual assessment of shape and size; and
- anthropometric/osteometric, i.e., based on instrumental measurements. The term "osteometric" applies to methods based on measurement of skeletal elements in particular.

* Corresponding author. Forensic Data Science Laboratory, Computer Science Department & Aston Institute for Forensic Linguistics, Aston University, Birmingham, UK.
E-mail address: geoff-morrison@forensic-evaluation.net (G.S. Morrison).

https://doi.org/10.1016/j.fsisyn.2021.100202
Received 27 July 2021; Received in revised form 23 September 2021; Accepted 23 September 2021
Available online 27 September 2021

Morphoscopic methods traditionally require considerable experience observing and understanding skeletal variation between individuals, populations, and age groups, and may be highly subjective in practice. Anthropometric methods are generally considered to be more objective, at least in the sense that intra- and inter-observer reliability is easier to assess. The most commonly used anthropometric measurements are point to point distances and angles. Some practitioners use a combination of morphoscopic and anthropometric methods.

It is common in forensic anthropology to draw inferences using statistical models applied to anthropometric data. A recently published book on the use of statistics and probability in forensic anthropology Obertová et al. [4], for instance, includes multiple chapters by different authors describing multiple statistical methods, including cluster analysis [5], logistic regression [6], and discriminant function analysis [7].

Use of classification models is common, and binary classification models have long been used to draw inferences with respect to biological sex, e.g., [8-13]. Commonly used models such as linear discriminant analysis, quadratic discriminant analysis, and logistic regression can output posterior probabilities, but in the forensic-anthropology literature a threshold is usually applied in order to obtain a classification.[1] For example, if the posterior probability for "male" is greater than 0.5 (or equivalently the posterior probability for "female" is less than 0.5) then the bone is classified as coming from a male, and if the posterior probability for "male" is less than 0.5 (or equivalently the posterior probability for "female" is greater than 0.5) then the bone is classified as coming from a female. In the forensic-anthropology literature, e.g., [14–16], there is evidence of some unease with this "fall-off-the-cliff" approach in which results with very different posterior probabilities, e.g., 0.51 and 0.99 are treated the same but results with very similar posterior probabilities, e.g., 0.49 and 0.51 are treated differently.

Galeta & Brůžek [7] reviews literature that expresses concern about a "zone of uncertainty", see Fig. 1. In this "zone of uncertainty" the posterior probability is relatively close to 0.5, and the probability that a bone will be misclassified is relatively high. In order to attempt to avoid misclassification, a procedure is adopted whereby the bone is not classified unless the posterior probability is relatively far from 0.5, e.g., if the posterior probability for "male" is greater than 0.95 then the remains are classified as male, and if the posterior probability for "male" is less than 0.05 then the remains are classified as female, but if the posterior probability for "male" is between 0.05 and 0.95 the remains are not classified as either male or female. In this example, the "zone of uncertainty" is between posterior probabilities of 0.05 and 0.95. Galeta & Brůžek [7] states that "It is a conservative approach, but it brings a high confidence of sex estimation at both the individual and the population level." The aim is to have a high correct-classification rate (a low classification-error rate) for the bones that are classified,[2] but this comes at the cost of not classifying some bones and in fact not drawing any inference about the sex of the latter bones. Non-classification can occur in a high proportion, even the majority, of cases. Galeta & Brůžek [7] discusses trade-off between correct-classification rate and proportion of cases not classified.

Bartholdy et al. [18] propose reporting the correct-classification rate corresponding to the posterior-probability value calculated for the bone of interest. They propose either calculating the correct-classification rate at the exact posterior-probability value obtained, or precalculating the correct-classification rate for a number of preselected posterior-probability threshold values, e.g., 0.8, 0.9, 0.95, then, once the posterior-probability value for the bone of interest is obtained, selecting the relevant precalculated result, i.e., if the exact posterior-probability value obtained is between 0.8 and 0.9, report the correct-classification rate that was precalculated excluding test results with posterior-probability values between 0.2 and 0.8, if the exact posterior-probability value obtained is between 0.9 and 0.95, report the correct-classification rate that was precalculated excluding test results with posterior-probability values between 0.1 and 0.9, etc. Bartholdy et al. [18] also suggests that results could be reported as "female", "probable female", "probable male", and "male" for posterior-probability ranges of, e.g., 0–0.2, 0.2–0.5, 0.5–0.8, and 0.8–1 respectively (see Fig. 1). Jerković et al. [17] propose the inverse solution of choosing a desired correct-classification rate and then finding the posterior-probability range that should be excluded in order to obtain this correct-classification rate.[3]

In the present paper, we propose what we believe is a simpler solution to the concerns expressed in the forensic-anthropology literature. We propose a move away from approaches in which the output is discretized into two or more bins, to an approach which makes direct use of continuously-valued output. Statistical models that can implement this approach are already widely used in forensic anthropology – all that is required to adopt this approach are minor modifications in the way those models are used and a change in the way practitioners and researchers think about the meaning of the output of the models. What we propose is implementation of the likelihood-ratio framework using relevant data, quantitative measurements, and statistical models.

We focus on explaining how to calculate likelihood ratios using linear discriminant analysis, quadratic discriminant analysis, and logistic regression models applied to osteometric data. For simplicity of exposition, we use data consisting of measurements made on a single skeletal element from each individual. The skeletal element we use is a humerus – humeri exhibit sexual dimorphism. The computer code for performing the calculations described in the present paper is provided at http://geoff-morrison.net/#LR_anthropology_2021. Parallel versions of the code are provided for Matlab, Python, and R.

2. Likelihood-ratio framework

Use of the likelihood-ratio framework is advocated by many who work in the area of forensic inference and statistics, e.g., Aitken et al. [19] with 31 authors/supporters, Morrison et al. [20] with 19 authors/supporters, and Morrison et al. [21] with 20 authors/supporters. Its use is also recommended in guidance documents issued by the following organizations:

[1] In the forensic-anthropology literature, the term "sectioning point" is often used rather than "threshold".

[2] Jerković et al. [17] claims that a 95% correct-classification rate "is the minimal level set by modern forensic and legal standards". We traced the publications that Jerković et al. [17] cited in support of this claim and the publications cited in those publications, but could find no support for the claim that this is a legal requirement. Nor could we find any evidence that it is a requirement of any standard on forensic science developed by a national or international standards-development organization.

[3] Note that if the data used for training and testing the statistical models were sampled from the same populations and the population distributions conformed to the assumptions of the models, then the expected value of the correct-classification rate would be predictable from the posterior-probability threshold and vice versa. If the posterior-probability threshold were τ, i.e., only data with posterior probabilities for male $p(H_M) \geq \tau$ and $p(H_M) \leq (1-\tau)$, or equivalently $p(H_F) \geq \tau$, were used for calculating the correct-classification rate κ, then the expected correct-classification rate would be $\kappa = (p(p(H_M) \geq \tau) + p(p(H_F) \geq \tau))/2 = (\tau + \tau)/2 = \tau$, e.g., if $\tau = 0.95$ then the expected value for $\kappa = 0.95$. Despite the difference in name, "posterior-probability threshold" versus "correct-classification rate", τ and κ represent the same underlying concept, with τ being the predicted value and κ being the empirically derived value. In practice τ and κ would usually differ because of violations of model assumptions, model overfitting, and/or sampling variability. With respect to sampling variability, keeping τ fixed but changing the sample used for training or the sample used for testing would usually result in a different value for κ. Separate sets of training and test data are used to assess the extent to which the model is useful.

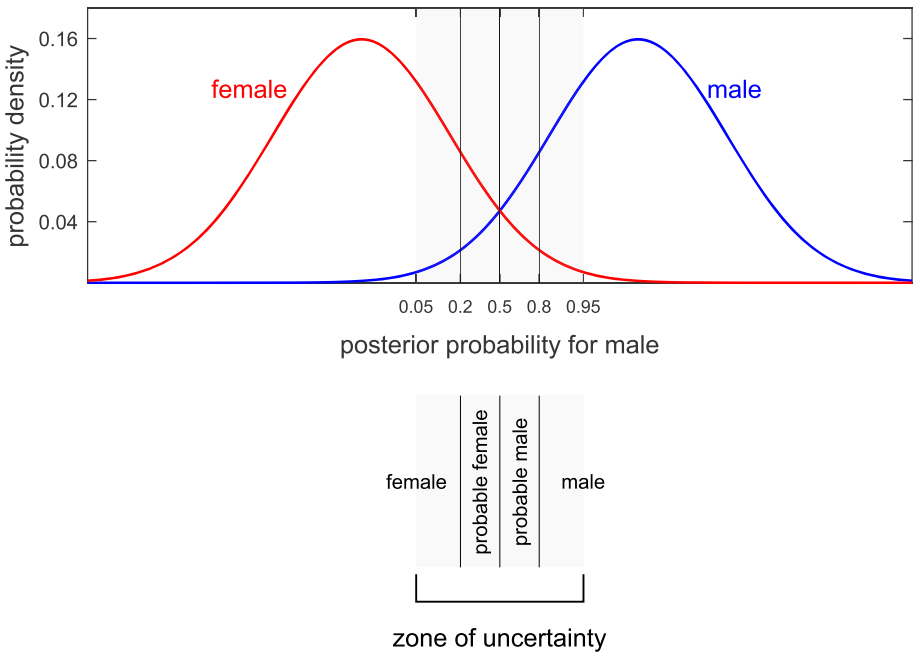

Fig. 1. Example (based on humeral-head-diameter data from [18]) of a univariate linear discriminant analysis model showing multiple threshold values at different posterior probabilities for the hypothesis that the osteometric measurement comes from a male. In this example, the prior probabilities for "male" versus "female" are equal. Also shown are a "zone of uncertainty" between posterior probabilities of 0.05 and 0.95, and verbal expressions corresponding to the posterior probability ranges 0–0.2, 0.2–0.5, 0.5–0.8, and 0.8–1 (the latter proposed in [18]).

- Association of Forensic Science Providers of the United Kingdom and of the Republic of Ireland (AFSP)[4] in 2009 [22].
- Royal Statistical Society (RSS)[5] in 2010 [23].
- European Network of Forensic Science Institutes (ENFSI)[6] in 2015 [24].
- National Institute of Forensic Science of the Australia New Zealand Policing Advisory Agency (NIFS ANZPAA)[7] in 2017 [25].
- American Statistical Association (ASA)[8] in 2019 [26].
- Forensic Science Regulator for England & Wales (FSR)[9] in 2021 [27].

Introductory texts on the likelihood-ratio framework include [28-35]. Publications advocating or describing application of the likelihood-ratio framework in forensic anthropology include [36-41].[10]

In the present paper, we do not attempt to provide a general introduction to the likelihood-ratio framework and arguments in favour of its use. Such information can be found in the references listed above. Instead, we focus on how to calculate likelihood ratios using the kinds of data and statistical models already familiar to practitioners and researchers in forensic anthropology. More complicated models can be used, and could potentially result in better performance, but for simplicity we focus on linear discriminant analysis, quadratic discriminant analysis, and logistic regression.[11]

For illustrative purposes, we use the humeral-measurement data from Bartholdy et al. [18]. The dataset contains measurements of maximum length, head diameter, and epicondylar breadth from the humeri of 36 males and 48 females. The dataset is small and the population does not reflect one that would be relevant for any modern forensic case, but it is a convenient dataset that will suffice to illustrate some statistical concepts. For univariate models we use the head-diameter measurements, and for bivariate models we use both head-diameter and epicondylar-breadth measurements.

The introductory literature on the likelihood-ratio framework tends to focus on what is often called "source attribution" or "individualization", e.g., situations in which a legal-decision maker wants to decide whether the bone in question comes from a particular individual or from some other individual randomly selected from a specified relevant population. Here, we focus on a simpler "classification" problem with

[4] http://www.afsp.org.uk/.
[5] https://rss.org.uk/.
[6] https://enfsi.eu/.
[7] https://www.anzpaa.org.au/forensic-science/nifs-home/.
[8] https://www.amstat.org/.
[9] https://www.gov.uk/government/organisations/forensic-science-regulator/.

[10] The likelihood-ratio framework for evaluation of forensic evidence should not be confused with likelihood-ratio tests used to assess difference in goodness of fit between competing models. Konigsberg et al. [42], for example, makes use of likelihood-ratio tests. Other references to likelihood ratios in that paper, e.g., "Taken as an evidentiary problem and assuming equal priors for male as for female within the population at large, the LR from the quadratic discriminant function is 1.997. This is found by calculating the [multivariate normal] density for obtaining 'Mr. Johnson's' measurements from the males and from the females …, averaging these densities across the two sexes, and dividing the male density by this average." (p. 80), are not likelihood ratios as understood in the likelihood-ratio framework. As defined in the quote, they are twice the posterior probability. The definition in the quote is equivalent to our Eq. (1) multiplied by two and assuming equal priors. The likelihood ratio corresponding to the value stated in the quote would actually be 666.

[11] Discriminant analysis assumes that the data from each class have Gaussian distributions, and linear discriminant analysis further assumes that the distributions from all classes have the same variance (in the univariate case) or the same covariance matrix (in the multivariate case). Histogram plots of the Bartholdy et al. [18] data reveal that these assumptions do not hold for epicondylar-breadth measurements: the female data appear to have a positive skew and the male data appear to be bimodal. Logistic regression is more robust to violations of these assumptions, but will not be robust to the bimodal distribution of the male data. Exploratory analysis of the data therefore suggest that none of linear discriminant analysis, quadratic discriminant analysis, or logistic regression are appropriate. Our purpose here, however, is simply to illustrate how to use these models, that are common in forensic anthropology, to calculate likelihood ratios. Whether these are good models to apply to these data, how well they perform when applied to these data, and the likelihood-ratio values that they output when applied to these data are not actually of concern. Use of linear discriminant analysis and logistic regression in the present paper also allows for direct comparison with their use in Bartholdy et al. [18] with the same dataset.

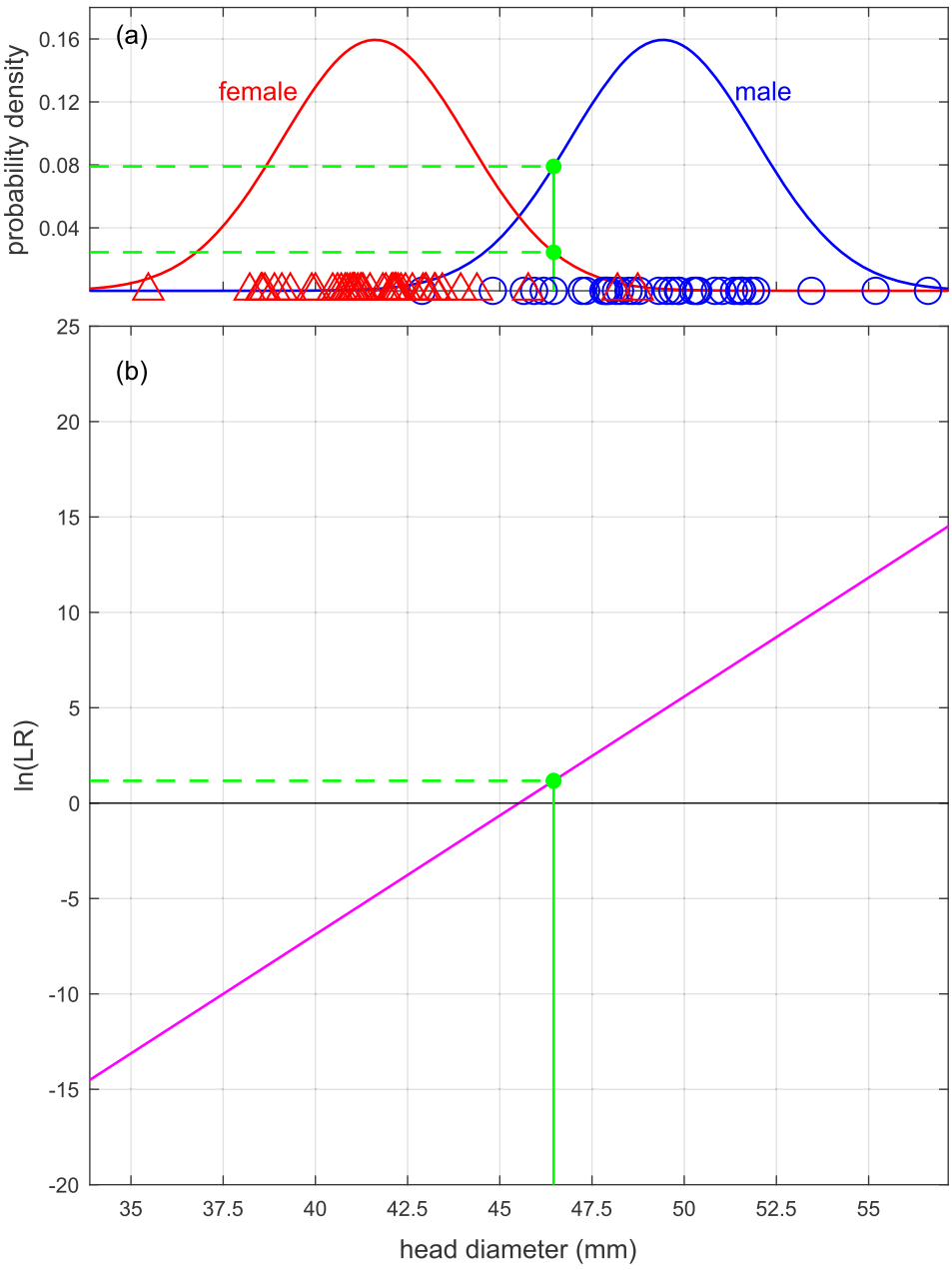

Fig. 2. Example (based on humeral-head-diameter data from [18]) of calculation of likelihood ratio using a univariate linear discriminant analysis model. (a): Calculation based on probability-density functions. (b): Calculation based on a linear equation.

only two mutually-exclusive classes, e.g., a situation in which a legal-decision maker's task is to decide whether the skeletal element in question comes from a male or from a female from the specified relevant population.

3. Calculating a likelihood ratio using linear discriminant analysis

Traditionally in forensic anthropology, linear discriminant analysis is used to calculate a posterior probability to which a threshold is then applied to make a classification. When first developed, without the aid of modern computers, calculations for linear discriminant analysis were laborious. Linear discriminant functions were therefore used ([43], [44]). For a two-class problem, multivariate data could be transformed to values on a univariate linear discriminant function, and, assuming equal priors, each test datum could then be classified according to whether it was closer to the centroid of one class or the other. A higher prior probability for one class, and concomitantly lower prior probability for the other, would shift the threshold on the linear discriminant function further from the centroid of the first class and closer to the centroid of the second class. The calculation of the linear discriminant function was laborious, but thereafter classifying test data was easy as it

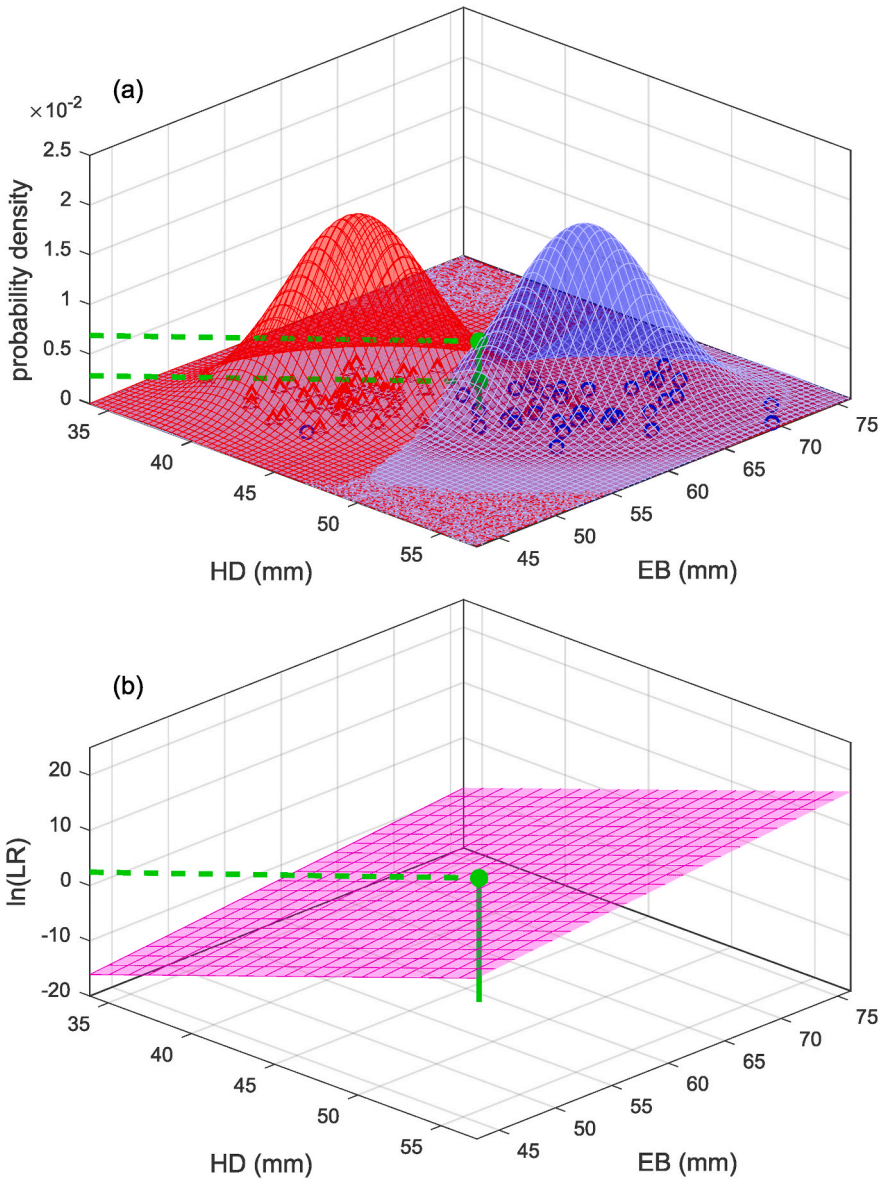

Fig. 3. Example (based on humeral head-diameter, HD, and epicondylar-breadth, EB, data from [18]) of calculation of likelihood ratio using a bivariate linear discriminant analysis model. (a): Calculation based on probability-density functions. (b): Calculation based on a linear equation.

did not require calculating the exact posterior probability for each new datum.

Using modern computers, the calculation of posterior probabilities (or of likelihoods) based on Gaussian distributions is trivial: all that is required is to enter training data into functions that calculate mean vectors and covariance matrices, then enter those statistics and the test data into functions that calculate probability densities. These functions are easily accessible in many programming languages and software packages. A posterior probability can be calculated as in Eq. (1), in which: H_M is the hypothesis that the humerus comes from a male in the relevant population; H_F is the hypothesis that the humerus comes from a female in the relevant population; $p(H_M|\boldsymbol{x}_Q)$ is the posterior probability that the "male" hypothesis H_M is true given the measurement vector $\boldsymbol{x}_Q$ from the bone in question; $f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the probability density (the like-

Table 1
Example likelihood-ratio values calculated using the same example measurement vector and different univariate and bivariate models.

	Univariate (head diameter)	Bivariate (head diameter, epicondylar breadth)
Linear discriminant analysis	3.23	2.44
Logistic regression	2.26	1.91
Quadratic discriminant analysis	4.22	2.64

lihood) of a Gaussian model with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ evaluated at vector $\boldsymbol{x}$; $\boldsymbol{\mu}_{\mathrm{M}}$ and $\boldsymbol{\mu}_{\mathrm{F}}$ are mean vectors calculated using a sample of data known to come from males in the relevant population and a sample of data known to come from females in the relevant population respectively; $\boldsymbol{\Sigma}$ is a covariance matrix calculated using data pooled from both the male and female samples[12]; $p(H_{\mathrm{M}})$ is the prior probability that the "male" hypothesis is true; and $p(H_{\mathrm{F}})$ is the prior probability that the "female" hypothesis H_{F} is true.

$$p(H_{\mathrm{M}}|\boldsymbol{x}_{\mathrm{Q}}) = \frac{f(\boldsymbol{x}_{\mathrm{Q}}|\boldsymbol{\mu}_{\mathrm{M}}, \boldsymbol{\Sigma})p(H_{\mathrm{M}})}{f(\boldsymbol{x}_{\mathrm{Q}}|\boldsymbol{\mu}_{\mathrm{M}}, \boldsymbol{\Sigma})p(H_{\mathrm{M}}) + f(\boldsymbol{x}_{\mathrm{Q}}|\boldsymbol{\mu}_{\mathrm{F}}, \boldsymbol{\Sigma})p(H_{\mathrm{F}})} \tag{1}$$

Since H_{M} and H_{F} are mutually exclusive and exhaustive, $p(H_{\mathrm{F}}) = 1 - p(H_{\mathrm{M}})$ and $p(H_{\mathrm{F}}|\boldsymbol{x}_{\mathrm{Q}}) = 1 - p(H_{\mathrm{M}}|\boldsymbol{x}_{\mathrm{Q}})$, and Eq. (1) can be rearranged to obtain Eq. (2), which is a version of the odds-form of Bayes' Theorem.

$$\frac{p(H_{\mathrm{M}}|\boldsymbol{x}_{\mathrm{Q}})}{p(H_{\mathrm{F}}|\boldsymbol{x}_{\mathrm{Q}})} = \frac{f(\boldsymbol{x}_{\mathrm{Q}}|\boldsymbol{\mu}_{\mathrm{M}}, \boldsymbol{\Sigma})}{f(\boldsymbol{x}_{\mathrm{Q}}|\boldsymbol{\mu}_{\mathrm{F}}, \boldsymbol{\Sigma})} \times \frac{p(H_{\mathrm{M}})}{p(H_{\mathrm{F}})} \tag{2}$$

posterior odds = likelihood ratio × prior odds

In the odds-form of Bayes' Theorem:

- the *prior odds* represent the legal-decision maker's belief as to the relative probability that the "male" hypothesis is true versus that the "female" hypothesis is true before they consider the forensic practitioner's statement of the strength of the evidence;
- the *likelihood ratio* is the forensic practitioner's statement of the strength of the evidence;
- and the *posterior odds* represent the legal-decision maker's belief as to the relative probability that the "male" hypothesis is true versus that the "female" hypothesis is true after they have considered the forensic practitioner's statement of the strength of the evidence.

The likelihood ratio therefore quantifies the amount by which, in light of the evidence, the legal-decision maker updates their belief with respect to the relative probabilities of the "male" and the "female" hypotheses. This assumes that the legal-decision maker is applying Bayes' Theorem and using the likelihood ratio provided by the forensic practitioner. These assumptions are adopted in order to explain the meaning of a likelihood ratio, not to describe how a legal-decision maker actually acts or to advise how a legal-decision maker should act.

For the likelihood-ratio value to be meaningful, one must also be satisfied that the data used for training the statistical models (e.g., the data used for calculating the mean vectors and the covariance matrix) are reasonably representative of the relevant population for the case.

The prior odds could be based on an estimate of the ratio of males to females in the relevant population, but could also depend on other evidence already presented in the case that has influenced the legal-decision maker's belief with respect to the relative probabilities of the two hypotheses.

In the likelihood-ratio framework, the task of the forensic practitioner is to assess and present the value of the likelihood ratio. The likelihood-ratio value can, in theory, be any number in the range 0 to $+\infty$ (the log-likelihood-ratio value can be any number in the range $-\infty$ to $+\infty$). The larger the number the greater the support it gives for the hypothesis in the numerator of the likelihood ratio (in this example, H_{M}), and the smaller the number the greater the support it gives for the hypothesis in the denominator of the likelihood ratio (in this example, H_{F}). If the likelihood-ratio value is 1 (the log-likelihood-ratio value is 0), it gives equal support for both hypotheses, and the posterior odds will be the same as the prior odds.

Assuming equal priors, $p(H_{\mathrm{M}}) = p(H_{\mathrm{F}})$, hence prior odds $p(H_{\mathrm{M}})/p(H_{\mathrm{F}}) = 1$, a "zone of uncertainty" based on posterior probability for male between 0.05 and 0.95 would correspond to likelihood-ratio values in the range 0.05/0.95 to 0.95/0.05 (= 1/19 to 19). Unlike an approach which does not draw any inference about the sex of bones with posterior probabilities within this "zone of uncertainty", likelihood ratios provide meaningful information both outside and within this range, and they do not suffer from a "fall-off-the-cliff" effect. Likelihood-ratio values of 2, 10, or 1/15, for example, provide information that a legal-decision maker could logically use to update their beliefs, and likelihood-ratio values of 18.9 and 19.1 will not be presented to legal-decision makers as if they had very different meanings.

Eq. (3) shows a univariate example of the calculation of a likelihood ratio $\Lambda(x)$, and Eq. (4) show a bivariate example of the calculation of a likelihood ratio $\Lambda(\boldsymbol{x})$. Fig. 2(a) shows a graphical representation of Eq. (3) in which the likelihood ratio for measurement scalar x is the height of the "male" curve relative to the height of the "female" curve, and Fig. 3(a) shows a graphical representation of Eq. (4) in which the likelihood ratio for measurement vector $\boldsymbol{x}$ is the height of the "male" surface relative to the height of the "female" surface. The values inserted into the equations and used to plot the figures are taken from the Bartholdy et al. [18] dataset. One measurement (from a male) was selected and used as $x = x_{\mathrm{Q}}$ in the univariate case and $\boldsymbol{x} = \boldsymbol{x}_{\mathrm{Q}}$ in the bivariate case (hereinafter we drop the Q subscript), and the remainder of the data were used to calculate the values for μ_{M}, μ_{F}, σ, $\boldsymbol{\mu}_{\mathrm{M}}$, $\boldsymbol{\mu}_{\mathrm{F}}$, and $\boldsymbol{\Sigma}$.

$$\begin{aligned}
\Lambda(x) &= \frac{f(x|\mu_{\mathrm{M}}, \sigma)}{f(x|\mu_{\mathrm{F}}, \sigma)} \\
&= \frac{f(x = 46.5 \mid \mu_{\mathrm{M}} = 49.4,\ \sigma = 2.50)}{f(x = 46.5 \mid \mu_{\mathrm{F}} = 41.6,\ \sigma = 2.50)} \\
&= \frac{0.0790}{0.0245} \\
&= 3.23
\end{aligned} \tag{3}$$

$$\begin{aligned}
\Lambda(\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|\boldsymbol{\mu}_{\mathrm{M}}, \boldsymbol{\Sigma})}{f(\boldsymbol{x}|\boldsymbol{\mu}_{\mathrm{F}}, \boldsymbol{\Sigma})} \\
&= \frac{f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \middle| \begin{bmatrix} \mu_{\mathrm{M},1} \\ \mu_{\mathrm{M},2} \end{bmatrix}, \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix}\right)}{f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \middle| \begin{bmatrix} \mu_{\mathrm{F},1} \\ \mu_{\mathrm{F},2} \end{bmatrix}, \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix}\right)} \\
&= \frac{f\left(\begin{bmatrix} 46.5 \\ 59.0 \end{bmatrix} \middle| \begin{bmatrix} 49.4 \\ 63.9 \end{bmatrix}, \begin{bmatrix} 6.26 & 4.84 \\ 4.84 & 15.8 \end{bmatrix}\right)}{f\left(\begin{bmatrix} 46.5 \\ 59.0 \end{bmatrix} \middle| \begin{bmatrix} 41.6 \\ 55.3 \end{bmatrix}, \begin{bmatrix} 6.26 & 4.84 \\ 4.84 & 15.8 \end{bmatrix}\right)} \\
&= \frac{0.00687}{0.00282} \\
&= 2.44
\end{aligned} \tag{4}$$

Table 1 collects the example likelihood-ratio values calculated using the same measurement vector $\boldsymbol{x}$ and all the different models presented in the present paper.

Before leaving linear discriminant analysis and moving on to logistic regression, in Eqs. 5–7 we show the derivation of the linear equation for the calculation of a likelihood ratio using linear discriminant analysis. For simplicity, we only show the derivation of the univariate equation: $y = a + bx$, in which y is the natural logarithm of the likelihood ratio, a is the intercept, b is the slope, and x is the head-diameter measurement made on the humerus.

[12] We used the formula for the unbiased estimate of the covariance matrix, i.e., dividing by $1 - n$ rather then by n (where n is the number of data point used to calculate the covariance matrix). We gave equal weight to each data point, i.e., we subtracted the class mean from the data in each class, pooled the data, and then calculated the covariance matrix.

$$
\begin{aligned}
y &= \ln(\Lambda(x)) = \ln\left(\frac{f(x|\mu_M,\sigma)}{f(x|\mu_F,\sigma)}\right)\\
&= \ln\left(\frac{\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\mu_M)^2}{-2\sigma^2}}}{\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\mu_F)^2}{-2\sigma^2}}}\right)\\
&= \ln\left(e^{\frac{(x-\mu_M)^2-(x-\mu_F)^2}{-2\sigma^2}}\right)\\
&= \frac{x^2+\mu_M^2-2x\mu_M-x^2-\mu_F^2+2x\mu_F}{-2\sigma^2}\\
&= \frac{-\mu_M^2+2x\mu_M+\mu_F^2-2x\mu_F}{2\sigma^2}\\
&= \frac{-\mu_M^2+\mu_F^2}{2\sigma^2}+\frac{\mu_M-\mu_F}{\sigma^2}x\\
&= a+bx
\end{aligned}
\tag{5}
$$

$$
b = \frac{\mu_M-\mu_F}{\sigma^2} \tag{6}
$$

$$
a = \frac{-\mu_M^2+\mu_F^2}{2\sigma^2} = -b\frac{\mu_M+\mu_F}{2} \tag{7}
$$

In Eqs. (8) and (9), we show a univariate example of the calculation of a likelihood ratio $\Lambda(x)$ given the same values as previously used in Eq. (3). Note that the final result in Eq. (9) is the same as the final result in Eq. (3). The same example is graphically represented in Fig. 2(b). Note that the straight line in Fig. 2(b) could be constructed by sweeping a probe along the x axis of Fig. 2(a) and at each point calculating the natural logarithm of the height of the "male" curve relative to the height of the "female" curve.

$$
\begin{aligned}
&y = a+bx\\
&= \frac{-\mu_M^2+\mu_F^2}{2\sigma^2}+\frac{\mu_M-\mu_F}{\sigma^2}x\\
&= \frac{-49.4^2+41.6^2}{2\times 2.50^2}+\frac{49.4-41.6}{2.50^2}\times 46.5\\
&= -56.8+1.25\times 46.5\\
&= 1.17
\end{aligned}
\tag{8}
$$

$$
\Lambda(x) = e^y = e^{1.17} = 3.23 \tag{9}
$$

The bivariate example is graphically represented in Fig. 3(b). Note that the plane in Fig. 3(b) could be constructed by sweeping a probe around the x_1-x_2 plane of Fig. 3(a) and at each point calculating the natural logarithm of the height of the "male" surface relative to the height of the "female" surface. The multivariate equation in general would be: $y = \beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_mx_m$, in which β_0 is the intercept and $\beta_1,\ldots,\beta_m$ are the slopes corresponding to the m dimensions of the data.

4. Calculating a likelihood ratio using logistic regression

Traditionally in forensic anthropology, logistic regression is used to calculate a posterior probability to which a threshold is then applied to make a classification. A posterior probability can be calculated as in Eq. (10), in which β_0 is an intercept and $\beta_1,\ldots,\beta_m$ are slopes corresponding to the m dimensions of the data. The values for $\beta_0,\ldots,\beta_m$ are calculated using an iterative algorithm. We do not describe the details of fitting logistic-regression models here, the interested reader is referred to texts such as [45] and [46]. For our calculations, we used the Newton iterative fitting algorithm with conjugate gradient ascent.

$$
p(H_M|\boldsymbol{x}) = \frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_mx_m)}} \tag{10}
$$

Since H_M and H_F are mutually exclusive and exhaustive, $p(H_F|\boldsymbol{x}) = 1-p(H_M|\boldsymbol{x})$, and Eq. (10) can be rearranged to obtain Eq. (11). Eq. (11) gives the logged posterior odds, and this is the form in which the model is actually fitted.

$$
\ln\left(\frac{p(H_M|\boldsymbol{x})}{p(H_F|\boldsymbol{x})}\right) = \beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_mx_m \tag{11}
$$

In order to use logistic regression to calculate a likelihood ratio, the data points in the training data should be weighted such that the two classes have the same weight; hence, $p(H_M) = p(H_F)$, the prior odds $p(H_M)/p(H_F) = 1$, and the posterior odds will equal the likelihood ratio (see Eq. (2)). Eqs. 12–15 repeat the same examples as for linear discriminant analysis above but with the coefficients values (a and b, and β_0, β_1, β_2) obtained using logistic regression. For parallelism with the linear equation derived for linear discriminant analysis in the previous section, the univariate example uses a and b for the intercept and slope. Note that the values for a and b in Eq. (12) are not the same as those obtained using linear discriminant analysis in Eq. (8). Figs. 4 and 5 show a graphical representation of the calculation of the likelihood ratio for the univariate and bivariate examples. Compare Figs. 4(b) and 5(b) with Figs. 2(b) and 3(b) respectively. In these examples, the slopes obtained using logistic regression are all shallower then the slopes obtained using linear discriminant analysis.

$$
\ln(\Lambda(x)) = y = a+bx = -42.3+0.928\times 46.5 = 0.815 \tag{12}
$$

$$
\Lambda(x) = e^y = e^{0.815} = 2.26 \tag{13}
$$

$$
\ln(\Lambda(\boldsymbol{x})) = y = \beta_0+\beta_1x_1+\beta_2x_2 = -42.9+0.809\times 46.5+0.102\times 59.0 = 0.648 \tag{14}
$$

$$
\Lambda(\boldsymbol{x}) = e^y = e^{0.648} = 1.91 \tag{15}
$$

Logistic regression is a discriminative model, not a generative model – it does not actually calculate the ratio of two likelihoods – but under ideal circumstances it would give the same results as linear discriminant analysis ([47] §4.4.5).[13] Because of its analogy with linear discriminant analysis, a generative model which actually calculates the ratio of two likelihoods, the output of logistic regression can be interpreted as a log likelihood ratio. Because it is not dependent on the assumptions of Gaussian distributions with the same covariance matrix, logistic regression is more robust than linear discriminant analysis when the data deviate from those assumptions. If the assumptions are met and the sample size is small; however, linear discriminant analysis will be less prone to overfit the training data.

5. Calculating a likelihood ratio using quadratic discriminant analysis

Quadratic discriminant analysis is the same as linear discriminant analysis, except that (in the present context) instead of using a single covariance matrix $\boldsymbol{\Sigma}$ calculated using data pooled from male and female samples, it uses two separate covariance matrices. $\boldsymbol{\Sigma}_M$ is calculated using data sampled from males and $\boldsymbol{\Sigma}_F$ is calculated using data sampled from females. Eq. (16) gives the quadratic-discriminant-analysis version of the odds-form of Bayes' Theorem, cf. Eq. (2).

$$
\frac{p(H_M|\boldsymbol{x}_Q)}{p(H_F|\boldsymbol{x}_Q)} = \frac{f(\boldsymbol{x}_Q|\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M)}{f(\boldsymbol{x}_Q|\boldsymbol{\mu}_F,\boldsymbol{\Sigma}_F)}\times\frac{p(H_M)}{p(H_F)} \tag{16}
$$

posterior odds = likelihood ratio × prior odds

Fig. 6 and Eq. (17) show the univariate example of the calculation of a likelihood ratio, and Fig. 7 and Eq. 18 show the bivariate example. Note that in Figs. 6(b) and 7(b) the mapping functions between x and

[13] A generative model is a model that estimates a probability density for the measurements.

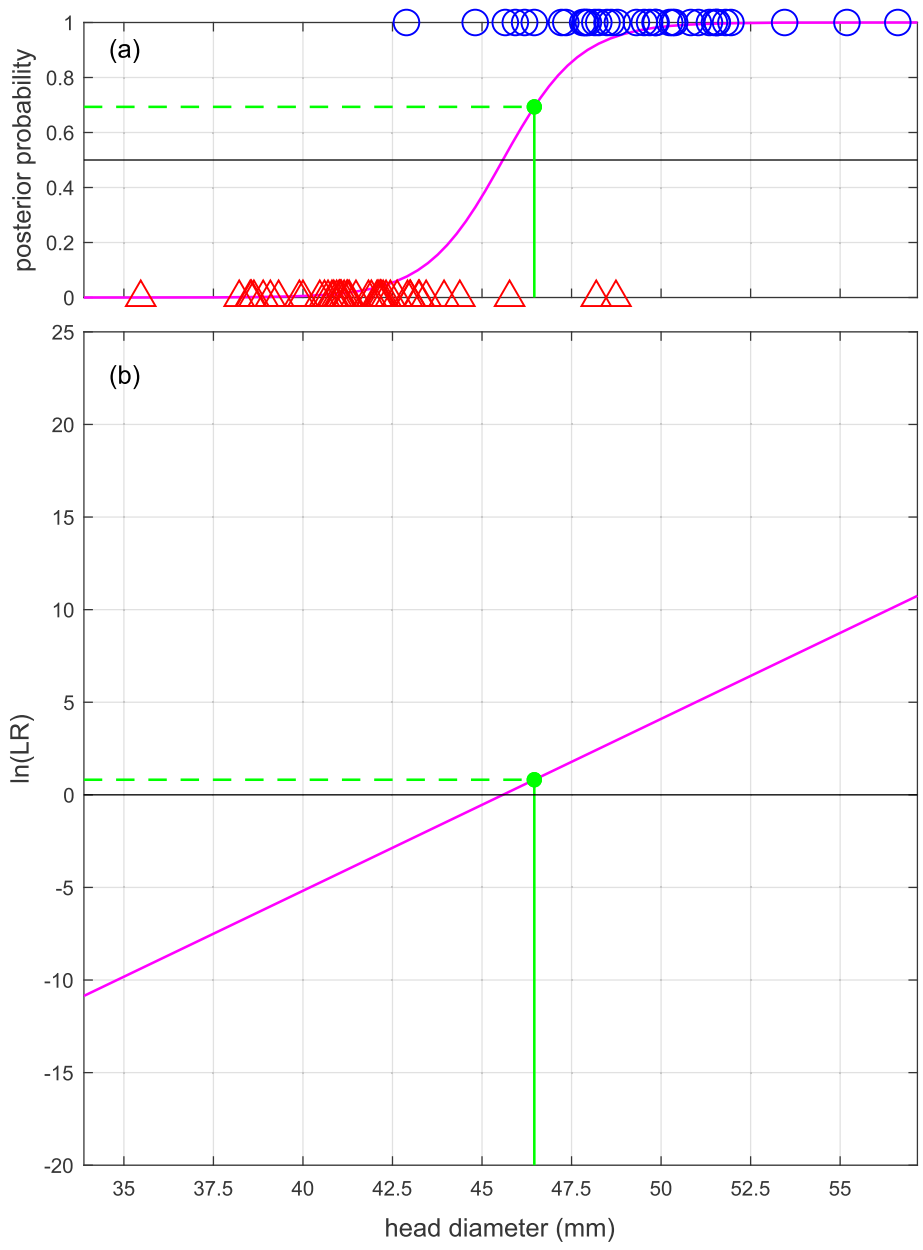

Fig. 4. Example (based on humeral-head-diameter data from [18]) of calculation of likelihood ratio using a univariate logistic regression model. Compare Fig. 4(b) with Fig. 2(b).

$\ln(\Lambda(\boldsymbol{x}))$ and between $\boldsymbol{x}$ and $\ln(\Lambda(\boldsymbol{x}))$ are not linear, they are a curve and a curved surface respectively.

$$\begin{aligned}\Lambda(x) &= \frac{f(x|\mu_{\mathrm{M}},\sigma_{\mathrm{M}})}{f(x|\mu_{\mathrm{F}},\sigma_{\mathrm{F}})}\\ &= \frac{f(x=46.5\mid \mu_{\mathrm{M}}=49.4,\ \sigma_{\mathrm{M}}=2.78)}{f(x=46.5\mid \mu_{\mathrm{F}}=41.6,\ \sigma_{\mathrm{F}}=2.31)}\\ &= \frac{0.0813}{0.0192}\\ &= 4.22\end{aligned} \tag{17}$$

$$\begin{aligned}\Lambda(\boldsymbol{x}) &= \frac{f(\boldsymbol{x}_{\mathrm{Q}}|\boldsymbol{\mu}_{\mathrm{M}},\boldsymbol{\Sigma}_{\mathrm{M}})}{f(\boldsymbol{x}_{\mathrm{Q}}|\boldsymbol{\mu}_{\mathrm{F}},\boldsymbol{\Sigma}_{\mathrm{F}})}\\ &= \frac{f\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}\middle|\middle|\begin{bmatrix}\mu_{\mathrm{M},1}\\\mu_{\mathrm{M},2}\end{bmatrix},\begin{bmatrix}\sigma_{\mathrm{M},1,1} & \sigma_{\mathrm{M},1,2}\\\sigma_{\mathrm{M},2,1} & \sigma_{\mathrm{M},2,2}\end{bmatrix}\right)}{f\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}\middle|\middle|\begin{bmatrix}\mu_{\mathrm{F},1}\\\mu_{\mathrm{F},2}\end{bmatrix},\begin{bmatrix}\sigma_{\mathrm{F},1,1} & \sigma_{\mathrm{F},1,2}\\\sigma_{\mathrm{F},2,1} & \sigma_{\mathrm{F},2,2}\end{bmatrix}\right)}\\ &= \frac{f\left(\begin{bmatrix}46.5\\59.0\end{bmatrix}\middle|\middle|\begin{bmatrix}49.4\\63.9\end{bmatrix},\begin{bmatrix}7.71 & 6.93\\6.93 & 23.2\end{bmatrix}\right)}{f\left(\begin{bmatrix}46.5\\59.0\end{bmatrix}\middle|\middle|\begin{bmatrix}41.6\\55.3\end{bmatrix},\begin{bmatrix}5.35 & 3.43\\3.43 & 10.7\end{bmatrix}\right)}\\ &= \frac{0.00680}{0.00258}\\ &= 2.64\end{aligned} \tag{18}$$

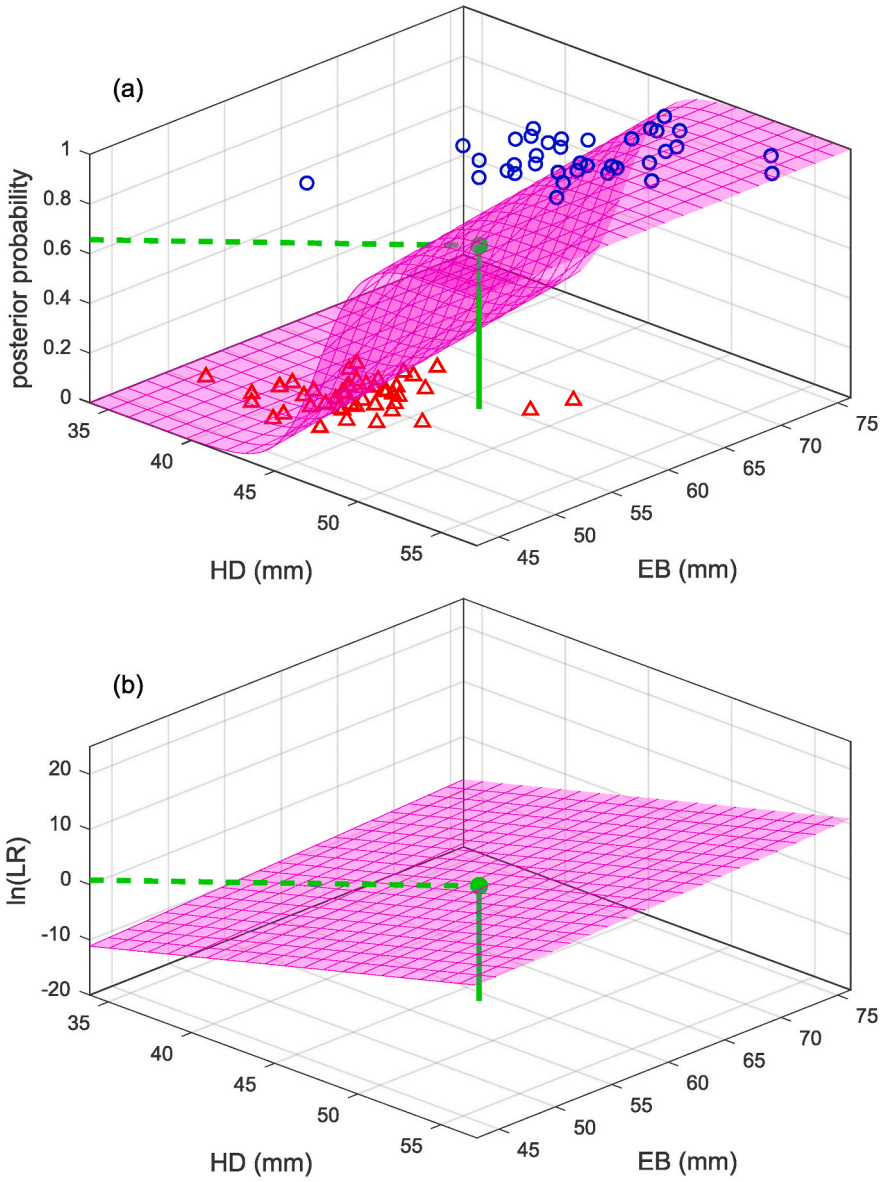

Fig. 5. Example (based on humeral head-diameter, HD, and epicondylar-breadth, EB, data from [18]) of calculation of likelihood ratio using a bivariate logistic regression model. Compare Fig. 5(b) with Fig. 3(b).

6. Validation of likelihood-ratio models

The performance of a model is assessed by:

1. Taking data that represent the relevant population for the case, that reflect conditions of the case, and for which the true class of each datum is known (e.g., each measurement vector is made on a humerus known to be from a male or know to be from a female from the population of interest);
2. Inputting each measurement vector into the model;
3. Then comparing the output of the model in response to each input with the known truth about the class of the corresponding input.

The test data must be separate from the data used to train the model, otherwise the results will be overly optimistic with respect to how well the model will perform when applied to previously unseen data, e.g., the measurements made on the humerus of questioned biological sex in the case.

Typically in the forensic-anthropology literature, the results are summarized using correct-classification rate, i.e., the proportion of all inputs that were correctly classified.[14] In the examples used in the present paper, the class of each input is either "male" or "female". In a classification framework, the class of each output would be either "male" or "female". If there is an imbalance in the number of "male" inputs and the number of "female" inputs in the validation data, the correct-classification rate can be separately calculated for each input class, then the mean over both classes calculated.

An alternative to correct-classification rate is classification-error rate, which is the proportion of inputs that were misclassified. This is equivalent to one minus the correct-classification rate.[15] The classification-error rate, E_{class}, with equal weighting for each class can be calculated as in Eq. (19), in which N_{M} and N_{F} are the number of inputs in the validation data known to be from males and the number of inputs in the validation data

[14] In the forensic-anthropology literature, correct-classification rate is usually expressed as a percentage. In the present paper, we express it as a proportion.

[15] If classification-error rate and correct-classification rate are expressed as a percentages, the classification-error rate is 100 minus the correct-classification rate.

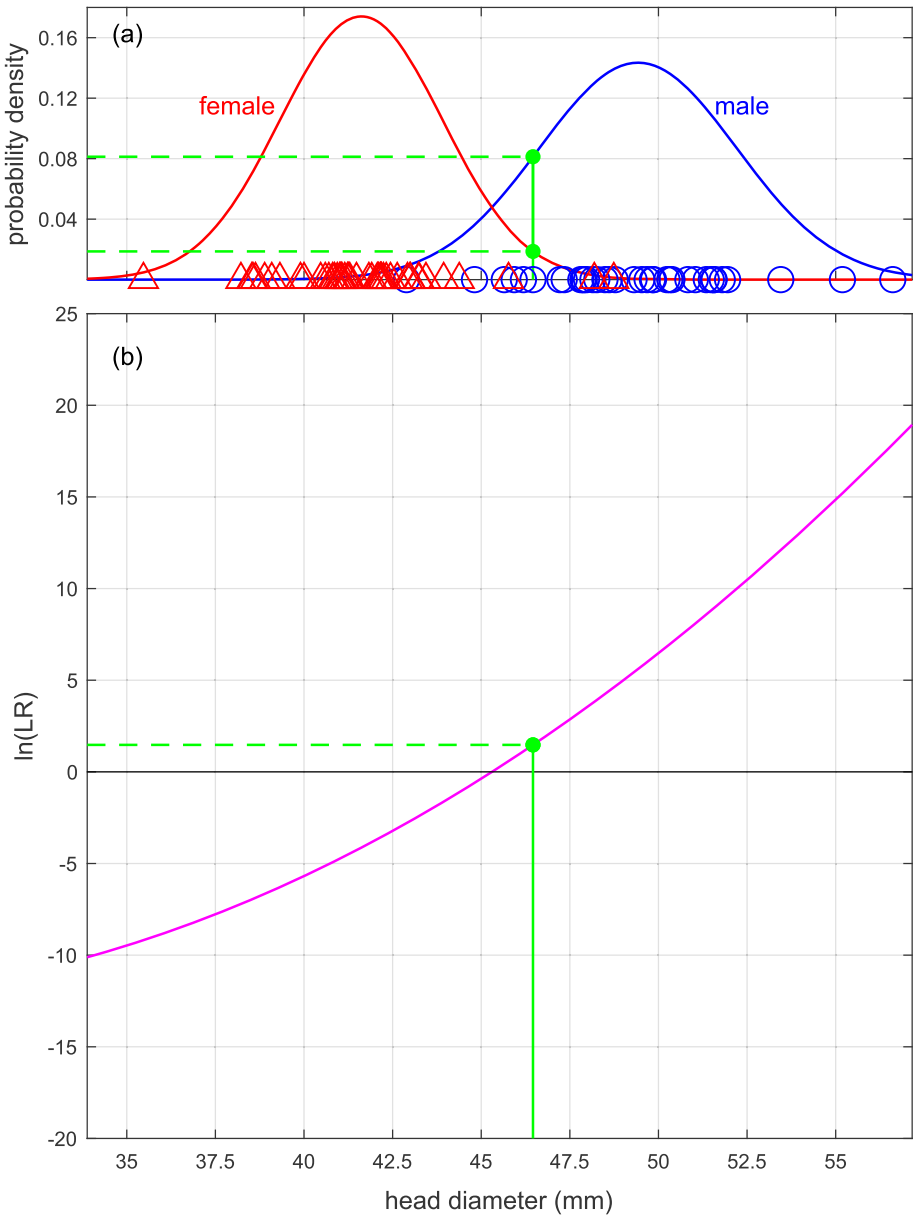

Fig. 6. Example (based on humeral-head-diameter data from [18]) of calculation of likelihood ratio using a univariate quadratic discriminant analysis model.

known to be from to be from females respectively, and Y_{M} and Y_{F} are classification outputs from the model in response to inputs known to be from males and inputs known to be from females respectively. In Eq. (19), a cost of 0 is assigned for a correct classification and a cost of 1 for an incorrect classification, the mean cost is calculated for inputs known to be from males and separately the mean cost is calculated for inputs known to be from females, then the mean of the latter two means is calculated. E_{class} is an average cost calculated over all the test data.

$$E_{\mathrm{class}} = \frac{1}{2}\left(\frac{1}{N_{\mathrm{M}}}\sum_{i}^{N_{\mathrm{M}}}\begin{pmatrix}0 \text{ if } Y_{\mathrm{M}_i} = \mathrm{M}\\ 1 \text{ if } Y_{\mathrm{M}_i} = \mathrm{F}\end{pmatrix} + \frac{1}{N_{\mathrm{F}}}\sum_{j}^{N_{\mathrm{F}}}\begin{pmatrix}0 \text{ if } Y_{\mathrm{F}_j} = \mathrm{F}\\ 1 \text{ if } Y_{\mathrm{F}_j} = \mathrm{M}\end{pmatrix}\right) \tag{19}$$

E_{class} is a number between 0 and 1 inclusive. Lower E_{class} values indicate better performance, i.e., fewer misclassifications. The expected E_{class} value for a model whose output was random would be 0.5. A model with an E_{class} value greater then 0.5 would be performing worse than chance.

In the likelihood-ratio framework, the output of the model is not a classification but a continuously-valued likelihood-ratio value. In our examples, which have H_{M} in the numerator and H_{F} in the denominator, the higher the likelihood-ratio value the greater the support for H_{M} relative to H_{F} and the lower the likelihood-ratio value the greater the support for H_{F} relative to H_{M}. If the input is from a male, the higher the likelihood-ratio value the greater the support for the correct hypothesis relative to the incorrect hypothesis. *Mutatis mutandis*, if the input is from a female, the lower the likelihood-ratio value the greater the support for the correct hypothesis relative to the incorrect hypothesis. Therefore, in order to assess the performance of a model that outputs likelihood ratios, we should not assign a cost of 0 or 1 based on classification, but rather a cost based on how good or how bad each likelihood-ratio values is:

- If we know the input was from a male we should assign a small cost value for a very large likelihood-ratio value, a larger cost value for a smaller likelihood-ratio value, and a much larger cost value for a very small likelihood-ratio value.
- *Mutatis mutandis*, if we know the input was from a female we should assign a small cost value for a very small likelihood-ratio value, a larger cost value for a larger likelihood-ratio value, and a much larger cost value for a very large likelihood-ratio value.

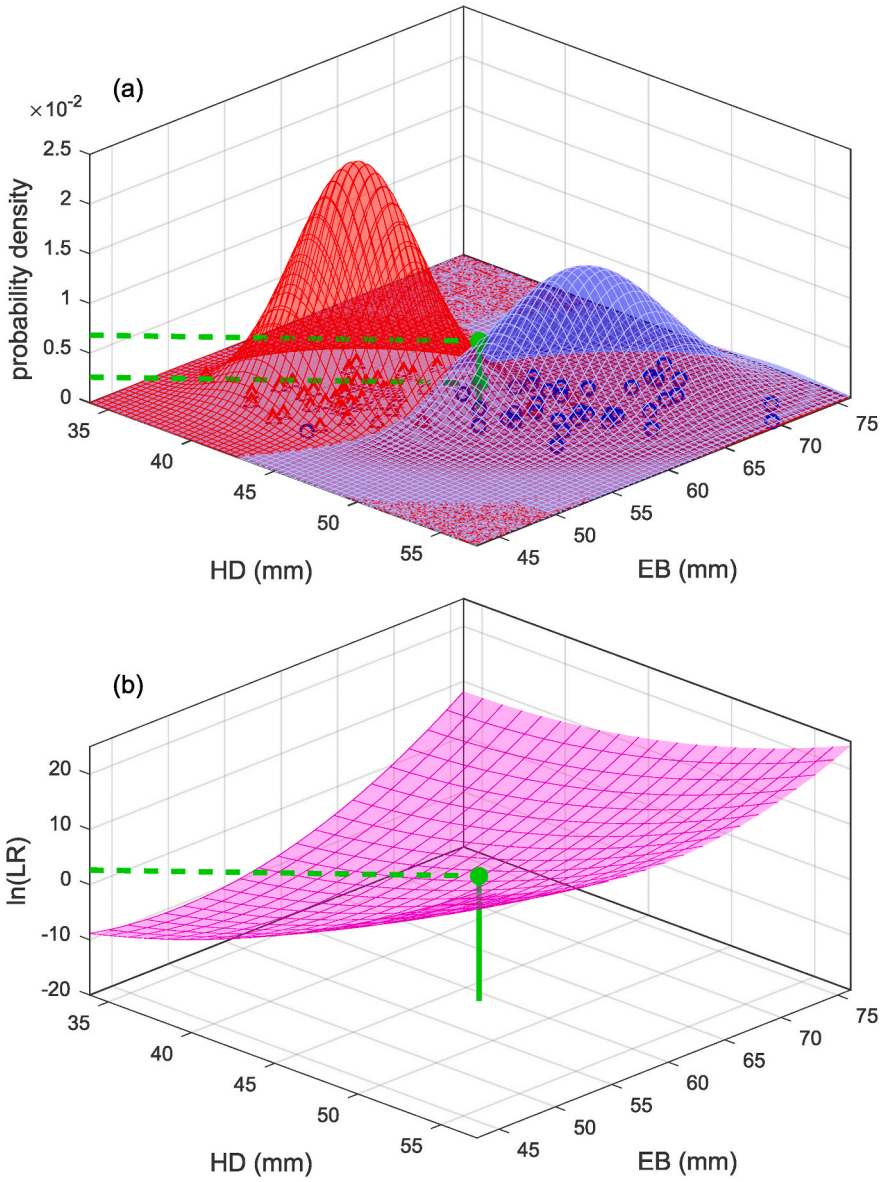

Fig. 7. Example (based on humeral head-diameter, HD, and epicondylar-breadth, EB, data from [18]) of calculation of likelihood ratio using a bivariate quadratic discriminant analysis model.

A commonly used metric in the forensic-inference-and-statistics literature (and especially in the forensic-voice-comparison literature [21]) is the log-likelihood-ratio cost, C_{llr} [48], see Eq. (20), in which Λ_{M} and Λ_{F} are likelihood-ratio outputs from the model in response to inputs known to be from males and inputs known to be from females respectively. The functions within the leftmost summation and rightmost summation of Eq. (20) are plotted in Fig. 8.

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{\text{M}}}\sum_{i}^{N_{\text{M}}} log_2\left(1+\frac{1}{\Lambda_{\text{M}_i}}\right) + \frac{1}{N_{\text{F}}}\sum_{j}^{N_{\text{F}}} log_2\left(1+\Lambda_{\text{F}_j}\right)\right) \tag{20}$$

C_{llr} is a number between 0 and $+\infty$. Lower C_{llr} values indicate better performance. A model that always output a likelihood ratio of 1 irrespective of the input would give no useful information: the posterior odds would always be the same as the prior odds. A model that gave no useful information would have a C_{llr} value of 1. Models that are miscalibrated can output likelihood ratios substantially larger than 1, but their performance can be improved by calibrating the system (see [49] for an introduction to this topic). Well calibrated systems will have C_{llr} values in the range 0 to ~1.

Returning to our univariate and bivariate examples, we validate the previously described models using leave-one-out cross validation, in which one measurement vector is held out, the remainder of the vectors are used to train the model, and the likelihood-ratio value is then calculated for the held-out vector. This is then repeated holding out each measurement vector in turn. This makes best use of the limited amount of data available while still avoiding training and testing on the same data. The resulting C_{llr} values are given in Table 2.

Based on the C_{llr} values in Table 2, the univariate models performed better than the bivariate models.[16] The simpler univariate linear models

[16] As mentioned in note 11, the epicondylar-breadth data violates the assumptions of all the models tested. Epicondylar breadth and head diameter were also highly correlated (Pearson's linear correlation coefficient $\rho = 0.794$). There may have been little additional useful information that the bivariate models could have exploited compared to their univariate counterparts, especially given the sampling variability associated with the small sample sizes. Univariate models based on epicondylar breadth had C_{llr} values in the range 0.5–0.6.

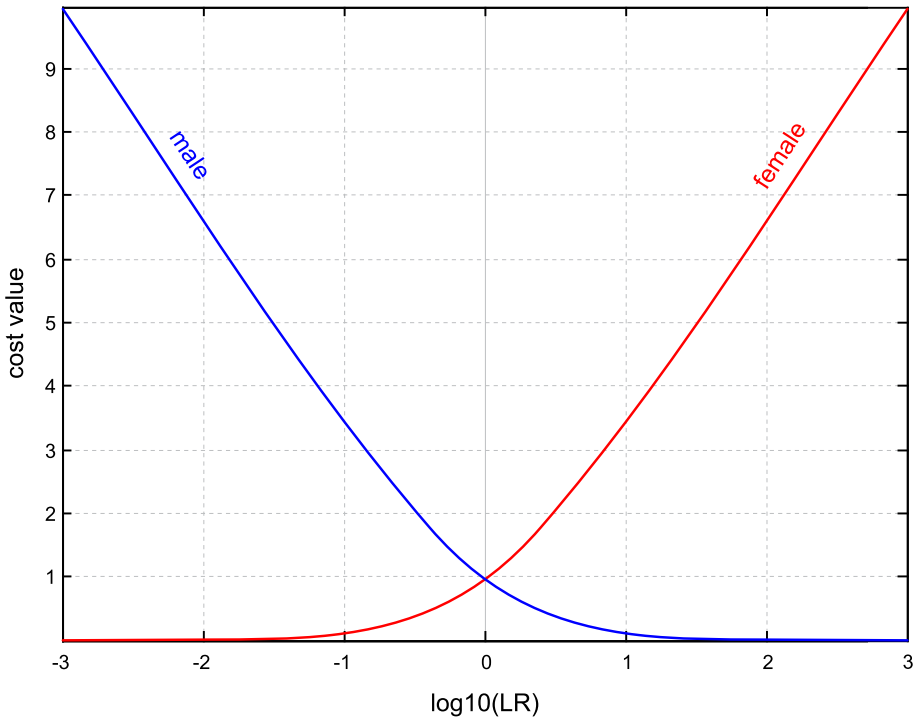

Fig. 8. Cost functions within the leftmost summation and rightmost summation of Eq. (20).

Table 2
values for different likelihood-ratio models applied to data from [18].

	Univariate (head diameter)	Bivariate (head diameter, epicondylar breadth)
Linear discriminant analysis	0.300	0.341
Logistic regression	0.306	0.349
Quadratic discriminant analysis	0.321	0.339

(linear discriminant analysis and logistic regression) also performed a little better than the more complex univariate quadratic discriminant analysis.

A graphical representation of likelihood-ratio validation results commonly used in the forensic-inference-and-statistics literature (and especially in the forensic-voice-comparison literature [21]) is a Tippett plot [50]. Tippett plots for the previously described likelihood-ratio models are given in Fig. 9. The likelihood-ratio value corresponding to each measurement vector is plotted as a dot, and straight lines are drawn between adjacent dots. In our examples, a Tippett plot displays the empirical cumulative distribution of all the likelihood-ratio values resulting from test data known to be from males, and the empirical cumulative distribution of all the likelihood-ratio values resulting from test data known to be from females. The empirical cumulative distributions are plotted so that for the curve rising to the right the value on the y axis is the proportion of male inputs resulting in likelihood-ratio values equal to or less than the value on the x axis, and for the curve rising to the left the value on the y axis is the proportion of female inputs resulting in likelihood-ratio values equal to or greater than the value on the x axis.

In general, the better the performance of the system that generated the likelihood-ratio results, the greater the separation between the "male" and "female" curves on the Tippett plots, and, concomitantly, the shallower the slopes of the curves. Given this, the results from quadratic discriminant analysis (shown in the bottom panels of Fig. 9) may appear to be better than the results from linear models (linear discriminant analysis and logistic regression shown in the top and middle panels), but the results from quadratic discriminant analysis also include some large-magnitude positive log-likelihood-ratio values for bones known to be from females. The results from the bivariate models (shown in the panels on the right) also include some large-magnitude positive log-likelihood-ratio values for bones known to be from females, and, in addition, some large-magnitude negative log-likelihood-ratio values for bones known to be from males. The extent of these likelihood-ratio results supporting contrary-to-fact hypotheses is less for the univariate linear models: univariate linear discriminant analysis and univariate logistic regression (shown in panels (a) and (c)). As already indicated by the C_{llr} values, the best results were obtained for the univariate linear models.

All models provide useful information, C_{llr} is substantially less than 1, and appear to give reasonably well-calibrated output – the curves in the Tippett plots cross relatively close to $\ln(LR) = 0$. For more complex models in which larger numbers of parameter values need to be estimated, it is usually necessary to calibrate their output using an explicit calibration model, see [51], [21], and [52].

Some of the models output likelihood-ratio values into the tens of thousands and even into the millions. These numbers are difficult to justify given the small sample sizes. To avoid complicating the present paper we do not address this issue here, but direct the interested reader to some solutions explored in [53].

Considering both C_{llr} and Tippett plots and the discussion above, given the Bartholdy et al. [18] dataset, the univariate logistic regression model appears to have resulted in the best performance. Note that it did not give the "best" results for the example feature vector (it did not give the largest likelihood-ratio value for this male feature vector), but it gave the best results averaged over all feature vectors. Given the small dataset, its lack of relevance for any modern forensic context, and the fact that the epicondylar-breadth data violate the assumptions of all the models tested, one should not draw any generalizations from any of the particular results presented here.

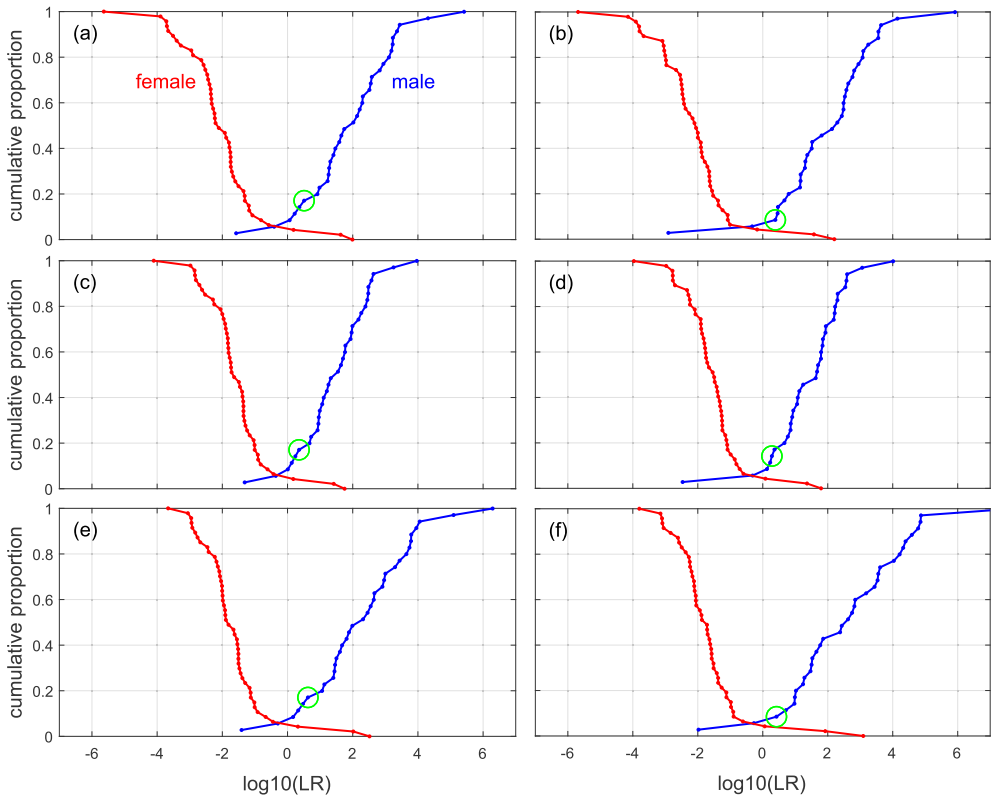

Fig. 9. Tippett plots for different likelihood-ratio models applied to data from [18]. (a) Univariate linear discriminant analysis. (b) Bivariate linear discriminant analysis. (c) Univariate logistic regression. (d) Bivariate logistic regression. (e) Univariate quadratic discriminant analysis. (f) Bivariate quadratic discriminant analysis. In each panel, the dot in the middle of the circle corresponds to the result from the example feature vector.

For other descriptions of both C_{llr} and Tippett plots see [54-57] and [21].

7. Conclusion

Use of the likelihood-ratio framework for evaluation of forensic evidence is advocated by many who work in the area of forensic inference and statistics, and in guidance documents issued by prominent organizations. So far, there has been little use of the likelihood-ratio framework in forensic anthropology, but, with respect to adoption of the likelihood-ratio framework, forensic anthropology has advantages over some other branches of forensic science: it is a branch of forensic science in which it is already common to draw inferences on the basis of relevant data, quantitative measurements, and statistical models. In the present paper, we explained how to calculate likelihood ratios using anthropometric data, and statistical models that are already commonly used in forensic anthropology: linear discriminant analysis, quadratic discriminant analysis, and logistic regression. We also explained how to empirically validate likelihood-ratio models. We hope that this will contribute to greater understanding and wider adoption of the likelihood-ratio framework in forensic-anthropology research and practice.

Disclaimer

All opinions expressed in the present paper are those of the authors, and, unless explicitly stated otherwise, should not be construed as representing the policies or positions of any organizations with which the authors are associated.

Author contributions

Geoffrey Stewart Morrison: Conceptualization, Writing – original draft, Writing – review & editing, Funding acquisition. Philip Weber: Investigation, Software, Visualization, Writing – review & editing. Nabanita Basu: Investigation, Software, Writing – review & editing. Roberto Puch-Solis: Software, Writing – review & editing. Patrick S. Randolph-Quinney: Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by Research England's Expanding Excellence in England Fund as part of funding for the Aston Institute for Forensic Linguistics 2019–2022.

References

[1] T.D. Stewart, Essentials of Forensic Anthropology, Charles C. Thomas, Springfield, IL, 1979.
[2] P.S. Randolph-Quinney, X. Mallett, S.M. Black, Anthropology, in: A. Jamieson, A. Moenssens (Eds.), Wiley Encylopedia of Forensic Science, Wiley, Chichester, UK, 2009, pp. 152–178.
[3] D.C. Dirkmaat, L.L Cabo, Forensic anthropology: embracing the new paradigm, in: D.C. Dirkmaat (Ed.), A Companion to Forensic Anthropology, Wiley Blackwell, Oxford, UK, 2012, pp. 3–40.
[4] Z. Obertová, A. Stewart, C. Cattaneo (Eds.), Statistics and Probability in Forensic Anthropology, Elsevier, London, UK, 2020, https://doi.org/10.1016/C2017-0-03461-4.
[5] Z. Obertová, A. Stewart, Probability distributions, hypothesis testing, and analysis, in: Z. Obertová, A. Stewart, C. Cattaneo (Eds.), Statistics and Probability in

Forensic Anthropology, Elsevier, London, UK, 2020, pp. 73–86, https://doi.org/10.1016/B978-0-12-815764-0.00011-3.
[6] E. Nikita, J.G. Gracía-Donad, P. Nikitas, Z. Obertová, E.F. Kranioti, Sex estimation using nonmetric variables: application of R functions, in: Z. Obertová, A. Stewart, C. Cattaneo (Eds.), Statistics and Probability in Forensic Anthropology, Elsevier, London, UK, 2020, pp. 139–154, https://doi.org/10.1016/B978-0-12-815764-0.00004-6.
[7] P. Galeta, J. Brůžek, Sex estimation using continuous variables: problems and principles of sex classification in the zone of uncertainty, in: Z. Obertová, A. Stewart, C. Cattaneo (Eds.), Statistics and Probability in Forensic Anthropology, Elsevier, London, UK, 2020, pp. 155–182, https://doi.org/10.1016/B978-0-12-815764-0.00016-2.
[8] J. Pons, The sexual diagnosis of isolated bones of the skeleton, Hum. Biol. 27 (1955) 12–21.
[9] E. Giles, O. Elliot, Sex determination by discriminant function analysis of crania, Am. J. Phys. Anthropol. 21 (1963) 53–68, https://doi.org/10.1002/ajpa.1330210108.
[10] S.R. Saunders, R.D. Hoppa, Sex allocation from long bone measurements using logistic regression, J. Can. Soc. Forensic. Sci. 30 (2) (1997) 49–60, https://doi.org/10.1080/00085030.1997.10757086.
[11] O. Ekizoglu, E. Inci, F.B. Palabiyik, I.O. Can, A. Er, M. Bozdag, I.E. Kacmaz, E. F. Kranioti, Sex estimation in a contemporary Turkish population based on CT scans of the calcaneus, Forensic Sci. Int. 279 (2017), https://doi.org/10.1016/j.forsciint.2017.07.038, 310e1–310e6.
[12] E. Nuzzolese, P. Randolph-Quinney, J. Randolph-Quinney, G. Di Vella, Geometric morphometric analysis of sexual dimorphism in the mandible from panoramic X-ray images, J. Forensic Odonto-Stomatology 37 (2) (2019) 35–44.
[13] M.A. Bidmos, A.A. Adebesin, P. Mazengenya, O.I. Olateju, O. Adegboy, Estimation of sex from metatarsals using discriminant function and logistic regression analyses, Aust. J. Forensic Sci. 53 (2021) 543–556, https://doi.org/10.1080/00450618.2019.1711180.
[14] P. Murail, J. Brůžek, J. Braga, A new approach to sexual diagnosis in past populations. practical adjustments from van Vark's procedure, Int. J. Osteoarchaeol. 9 (1999) 39–53, https://doi.org/10.1002/(SICI)1099-1212(199901/02)9:1<39::AID-OA458>3.0.CO;2-V.
[15] J. Brůžek, F. Santos, B. Dutailly, P. Murail, E. Cunha, Validation and reliability of the sex estimation of the human os coxae using freely available DSP2 software for bioarchaeology and forensic anthropology, Am. J. Phys. Anthropol. 164 (2017) 440–449, https://doi.org/10.1002/ajpa.23282.
[16] M. Hora, V. Sládek, Population specificity of sex estimation from vertebrae, Forensic Sci. Int. 291 (2018) 279, https://doi.org/10.1016/j.forsciint.2018.08.015, e1–279.e12.
[17] I. Jerković, Ž. Bašić, Š. Anđelinović, I. Kružić, Adjusting posterior probabilities to meet predefined accuracy criteria: a proposal for a novel approach to osteometric sex estimation, Forensic Sci. Int. 311 (2020), https://doi.org/10.1016/j.forsciint.2020.110273 article 110273.
[18] B.P. Bartholdy, E. Sandoval, M.L.P. Hoogland, S.A. Schrader, Getting rid of dichotomous sex estimations: why logistic regression should be preferred over discriminant function analysis, J. Forensic Sci. 65 (2020) 1685–1691, https://doi.org/10.1111/1556-4029.14482, 2020.
[19] C.G.G. Aitken, C.E.H. Berger, J.S. Buckleton, C. Champod, J.M. Curran, A.P. Dawid, I.W. Evett, P. Gill, J. González-Rodríguez, G. Jackson, A. Kloosterman, T. Lovelock, D. Lucy, P. Margot, L. McKenna, D. Meuwly, C. Neumann, N. Nic Daéid, A. Nordgaard, R. Puch-Solis, B. Rasmusson, M. Redmayne, P. Roberts, B. Robertson, C. Roux, M.J. Sjerps, F. Taroni, T. Tjin-A-Tsoi, G.A. Vignaux, S. M. Willis, G. Zadora, Expressing evaluative opinions: a position statement, Sci. Justice 51 (2011) 1–2, https://doi.org/10.1016/j.scijus.2011.01.002.
[20] G.S. Morrison, D.H. Kaye, D.J. Balding, D. Taylor, P. Dawid, C.G.G. Aitken, S. Gittelson, G. Zadora, B. Robertson, S.M. Willis, S. Pope, M. Neil, K.A. Martire, A. Hepler, R.D. Gill, A. Jamieson, J. de Zoete, R.B. Ostrum, A. Caliebe, A comment on the PCAST report: skip the "match"/"non-match" stage, Forensic Sci. Int. 272 (2017), https://doi.org/10.1016/j.forsciint.2016.10.018 e7–e9.
[21] G.S. Morrison, E. Enzinger, V. Hughes, M. Jessen, D. Meuwly, C. Neumann, S. Planting, W.C. Thompson, D. van der Vloed, R.J.F. Ypma, C. Zhang, A. Anonymous, B. Anonymous, Consensus on validation of forensic voice comparison, Sci. Justice 61 (2021) 229–309, https://doi.org/10.1016/j.scijus.2021.02.002.
[22] Association of Forensic Science Providers, Standards for the formulation of evaluative forensic science expert opinion, Sci. Justice 49 (2009) 161–164, https://doi.org/10.1016/j.scijus.2009.07.004.
[23] C.G.G. Aitken, P. Roberts, G. Jackson, Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, Royal Statistical Society, London, UK, 2010. https://rss.org.uk/news-publication/publications/law-guides/.
[24] S.M. Willis, L. McKenna, S. McDermott, G. O'Donell, A. Barrett, A. Rasmusson, A. Nordgaard, C.E.H. Berger, M.J. Sjerps, J.J. Lucena-Molina, G. Zadora, C.G. G. Aitken, L. Lunt, C. Champod, A. Biedermann, T.N. Hicks, F. Taroni, ENFSI Guideline for Evaluative Reporting in Forensic Science, European Network of Forensic Science Institutes, 2015. http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.
[25] K. Ballantyne, J. Bunford, B. Found, D. Neville, D. Taylor, G. Wevers, D. Catoggio, An Introductory Guide to Evaluative Reporting, National Institute of Forensic Science of the Australia New Zealand Policing Advisory Agency, 2017. http://www.anzpaa.org.au/forensic-science/our-work/projects/evaluative-reporting.
[26] K. Kafadar, H. Stern, M. Cuellar, J. Curran, M. Lancaster, C. Neumann, C. Saunders, B. Weir, S. Zabell, American Statistical Association Position on Statistical Statements for Forensic Evidence, American Statistical Association, 2019. https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf.
[27] Forensic Science Regulator, Codes of Practice and Conduct: Development of Evaluative Opinions (FSR-C-118 Issue 1), Forensic Science Regulator Birmingham, UK, 2021. https://www.gov.uk/government/publications/development-of-evaluative-opinions.
[28] D. Lucy, Introduction to Statistics for Forensic Scientists, Wiley, Chichester UK, 2005.
[29] G. Zadora, A. Martyna, D. Ramos, C.G.G. Aitken, Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data, Wiley, Chichester, UK, 2014, https://doi.org/10.1002/9781118763155.
[30] D.J. Balding, C. Steele, Weight-of-Evidence for Forensic DNA Profiles, second ed., Wiley, Chichester, UK, 2015 https://doi.org/10.1002/9781118814512.
[31] C. Adam, Forensic Evidence in Court: Evaluation and Scientific Opinion, Wiley, Chichester, UK, 2016, https://doi.org/10.1002/9781119054443.
[32] J.S. Buckleton, J.A. Bright, D. Taylor (Eds.), Forensic DNA Evidence Interpretation, second ed., CRC, Boca Raton, FL, 2016.
[33] B. Robertson, G.A. Vignaux, C.E.H. Berger, Interpreting Evidence: Evaluating Forensic Science in the Courtroom, second ed., Wiley, Chichester, UK, 2016 https://doi.org/10.1002/9781118492475.
[34] G.S. Morrison, E. Enzinger, C. Zhang, Forensic speech science, in: I. Freckelton, H. Selby (Eds.), Expert Evidence, Thomson Reuters, Sydney, Australia, 2018 ch. 99.
[35] C.G.G. Aitken, F. Taroni, S. Bozza, Statistics and the Evaluation of Evidence for Forensic Scientists, third ed., Wiley, Chichester, UK, 2021 https://doi.org/10.1002/9781119245438.
[36] H.H. de Boer, S. Blau, T. Delabarde, L. Hackman, The role of forensic anthropology in disaster victim identification (DVI): recent developments and future prospects, Forensic Science Research 4 (2019) 303–315, https://doi.org/10.1080/20961790.2018.1480460.
[37] H.H. de Boer, M. van Wijk, C.E.H. Berger, Communicating evidence with a focus on the use of Bayes' theorem, in: Z. Obertová, A. Stewart, C. Cattaneo (Eds.), Statistics and Probability in Forensic Anthropology, Elsevier, London, UK, 2020, pp. 331–340, https://doi.org/10.1016/B978-0-12-815764-0.00034-4.
[38] C.E.H. Berger, H.H. de Boer, M. van Wijk, Use of Bayes' Theorem in data analysis and interpretation, in: Z. Obertová, A. Stewart, C. Cattaneo (Eds.), Statistics and Probability in Forensic Anthropology, Elsevier, London, UK, 2020, pp. 125–135, https://doi.org/10.1016/B978-0-12-815764-0.00014-9.
[39] C.E.H. Berger, M. van Wijk, H.H. de Boer, Bayesian inference in personal identification, in: Z. Obertová, A. Stewart, C. Cattaneo (Eds.), Statistics and Probability in Forensic Anthropology, Elsevier, London, UK, 2020, pp. 301–312, https://doi.org/10.1016/B978-0-12-815764-0.00006-X.
[40] R. Verma, K. Krishan, D. Rani, A. Kumar, V. Sharma, Stature estimation in forensic examinations using regression analysis: a likelihood ratio perspective, Forensic Sci. Int.: Report 2 (2020), https://doi.org/10.1016/j.fsir.2020.100069 article 100069.
[41] R. Verma, K. Krishan, D. Rani, A. Kumar, V. Sharma, R. Shreshtha, T. Kanchan, Estimation of sex in forensic examinations using logistic regression and likelihood ratios, Forensic Sci. Int.: Report 2 (2020), https://doi.org/10.1016/j.fsir.2020.100118 article 100118.
[42] L.W. Konigsberg, B.F. Algee-Hewitt, D.W. Steadman, Estimation and evidence in forensic anthropology: sex and race, Am. J. Phys. Anthropol. 139 (2009) 77–90, https://doi.org/10.1002/ajpa.20934.
[43] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 179–188, https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.
[44] W.R. Klecka, Discriminant Analysis, Sage, Beverly Hills, CA, 1980.
[45] S. Menard, Logistic Regression: from Introductory to Advanced Concepts and Applications, Sage, Thousand Oaks, CA, 2010, https://doi.org/10.4135/9781483348964.
[46] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, third ed., Wiley, Hoboken, NJ, 2013 https://doi.org/10.1002/9781118548387.
[47] T. Hastie, R. Tibshirani, J. Freidman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, second ed., Springer, New York, 2009 https://doi.org/10.1007/978-0-387-84858-7.
[48] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Comput. Speech Lang 20 (2006) 230–275, https://doi.org/10.1016/j.csl.2005.08.001.
[49] G.S. Morrison, Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio, Aust. J. Forensic Sci. 45 (2013) 173–197, https://doi.org/10.1080/00450618.2012.733025.
[50] D. Meuwly, Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique, Doctoral dissertation, University of Lausanne, 2001.
[51] G.S. Morrison, E. Enzinger, D. Ramos, J. González-Rodríguez, A. Lozano-Díez, Statistical models in forensic voice comparison, in: D.L. Banks, K. Kafadar, D. H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, CRC, Boca Raton, FL, 2020, pp. 451–497, https://doi.org/10.1201/9780367527709.
[52] G.S. Morrison, In the context of forensic casework, are there meaningful metrics of the degree of calibration? Forensic Sci. Int.: Synergy 3 (2021) https://doi.org/10.1016/j.fsisyn.2021.100157 article 100157.
[53] G.S. Morrison, N. Poh, Avoiding overstating the strength of forensic evidence: shrunk likelihood ratios/Bayes factors, Sci. Justice 58 (2018) 200–218, https://doi.org/10.1016/j.scijus.2017.12.005.
[54] J. González-Rodríguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-García, Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Trans. Audio Speech Lang. Process. 15 (2007) 2104–2115, https://doi.org/10.1109/TASL.2007.902747.

[55] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, T. Niemi, Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition, Including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises, European Network of Forensic Science Institutes, 2015. http://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf.

[56] G.S. Morrison, E. Enzinger, Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction, Speech Commun. 85 (2016) 119–126, https://doi.org/10.1016/j.specom.2016.07.006.

[57] D. Meuwly, D. Ramos, R. Haraksim, A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic Sci. Int. 276 (2017) 142–153, https://doi.org/10.1016/j.forsciint.2016.03.048.